



UPGRADE is the European Journal for the Informatics Professional, published bimonthly at <http://www.upgrade-cepis.org/>

UPGRADE is the anchor point for UPENET (UPGRADE European Network), the network of CEPIS member societies' publications, that currently includes the following ones:

- **Mondo Digitale**, digital journal from the Italian CEPIS society AICA
- **Novática**, journal from the Spanish CEPIS society ATI
- **OCG Journal**, journal from the Austrian CEPIS society OCG
- **Pliroforiki**, journal from the Cyprus CEPIS society CCS
- **Pro Dialog**, journal from the Polish CEPIS society PTI-PIPS

Publisher

UPGRADE is published on behalf of CEPIS (Council of European Professional Informatics Societies, <http://www.cepis.org/>) by **Novática** (<http://www.ati.es/novatica/>), journal of the Spanish CEPIS society ATI (*Asociación de Técnicos de Informática*, <http://www.ati.es/>)

UPGRADE monographs are also published in Spanish (full version printed; summary, abstracts and some articles online) by **Novática**, and in Italian (summary, abstracts and some articles online) by the Italian CEPIS society ALSI (*Associazione nazionale Laureati in Scienze dell'informazione e Informatica*, <http://www.alsi.it/>) and the Italian IT portal **Tecnoteca** (<http://www.tecnoteca.it/>)

UPGRADE was created in October 2000 by CEPIS and was first published by **Novática** and **INFORMATIK/INFORMATIQUE**, bimonthly journal of SVI/FSI (Swiss Federation of Professional Informatics Societies, <http://www.svifsi.ch/>)

Editorial Team

Chief Editor: Rafael Fernández Calvo, Spain, rfcvalvo@ati.es
Associate Editors:
François Louis Nicolet, Switzerland, nicolet@acm.org
Roberto Carniel, Italy, carniel@dgt.uniud.it
Zakaria Maamar, Arab Emirates, Zakaria.Maamar@zu.ac.ae
Soraya Kouadri Mostéfaoui, Switzerland, soraya.kouadrimostefaoui@unifr.ch

Editorial Board

Prof. Wolfried Stucky, Former President of CEPIS
Prof. Nello Scarabottolo, CEPIS Vice President
Fernando Piera Gómez and
Rafael Fernández Calvo, ATI (Spain)
François Louis Nicolet, SI (Switzerland)
Roberto Carniel, ALSI – Tecnoteca (Italy)

UPENET Advisory Board

Franco Filippazzi (Mondo Digitale, Italy)
Rafael Fernández Calvo (Novática, Spain)
Veith Risak (OCG Journal, Austria)
Panicos Masouras (Pliroforiki, Cyprus)
Andrzej Marciniak (Pro Dialog, Poland)

English Editors: Mike Andersson, Richard Butchart, David Cash, Arthur Cook, Tracey Darch, Laura Davies, Nick Dunn, Rodney Fennemore, Hilary Green, Roger Harris, Michael Hird, Jim Holder, Alasdair MacLeod, Pat Moody, Adam David Moss, Phil Parkin, Brian Robson

Cover page designed by Antonio Crespo Foix, © ATI 2005

Layout Design: François Louis Nicolet

Composition: Jorge Llácer-Gil de Ramales

Editorial correspondence: Rafael Fernández Calvo rfcvalvo@ati.es

Advertising correspondence: novatica@ati.es

UPGRADE Newsletter available at

<http://www.upgrade-cepis.org/pages/editinfo.html#newsletter>

Copyright

© Novática 2005 (for the monograph and the cover page)

© CEPIS 2005 (for the sections MOSAIC and UPENET)

All rights reserved. Abstracting is permitted with credit to the source. For copying, reprint, or republication permission, contact the Editorial Team

The opinions expressed by the authors are their exclusive responsibility

ISSN 1684-5285

Monograph of next issue (February 2006):

Key Success Factors in Software Engineering
(The full schedule of UPGRADE is available at our website)

Monograph: The Semantic Web (published jointly with Novática*)

Guest Editors: *Luis Sánchez-Fernández, Michael Sintek, and Stefan Decker*

- 2 Presentation. The Semantic Web or The Next Web Wave – *Luis Sánchez-Fernández, Michael Sintek, and Stefan Decker*
- 5 The Semantic Web: Fundamentals and A Brief State-of-the-Art – *Luis Sánchez-Fernández and Norberto Fernández-García*
- 12 Leveraging Metadata Creation by Annotation for The Semantic Web – *Siegfried Handschuh*
- 19 The Quest for Information Retrieval on The Semantic Web – *David Vallet-Weadon, Miriam Fernández-Sánchez, and Pablo Castells-Azpilicuet*
- 24 Functional RuleML: From Horn Logic with Equality to Lambda Calculus – *Harold Boley*
- 30 Towards Semantic Desktop Wikis – *Malte Kiesel and Leo Sauerermann*
- 35 Towards Semantically-Interlinked Online Communities – *Uldis Bojars, John G. Breslin, Andreas Harth, and Stefan Decker*
- 41 A Semantic Search Engine for the International Relation Sector – *Luis Rodrigo-Aguado, V. Richard Benjamins, Jesús Contreras-Cino, Diego-Javier Patón-Villahermosa, David Navarro-Arno, Robert Salla-Figuerol, Mercedes Blázquez-Cívico, Pilar Tena-García, and Isabel Martos-Laborde*
- 48 Semantic Search in Digital Image Archives: A Case Study – *Julio Villena-Román, José-Carlos González-Cristóbal, Cristina Moreno-García, and José- Luis Martínez-Fernández.*
- 55 Configuring e-Government Services Using Ontologies – *Dimitris Apostolou, Ljiljana Stojanovic, Tomás Pariente-Lobo, Joan Battle-Montserrat, and Andreas E. Papadakis*

UPENET (UPGRADE European Network)

- 63 From **Novática** (ATI, Spain)
ICT for Education
An Initiative for Educational Modernization: The Ponte dos Brozos Project – *Simón Neira-Dueñas and Felipe Gómez-Pallete Rivas*
- 71 From **Pro Dialog** (PIPS, Poland)
ICT for Education
On The Superiority of Internet-Based Mass Enrolment to High Schools over Traditional – *Andrzej P. Urbanski*

* This monograph will be also published in Spanish (full version printed; summary, abstracts, and some articles online) by **Novática**, journal of the Spanish CEPIS society ATI (*Asociación de Técnicos de Informática*) at <http://www.ati.es/novatica/>, and in Italian (online edition only, containing summary, abstracts, and some articles) by the Italian CEPIS society ALSI (*Associazione nazionale Laureati in Scienze dell'informazione e Informatica*) and the Italian IT portal **Tecnoteca** at <http://www.tecnoteca.it/>.

Semantic Search in Digital Image Archives: A Case Study

Julio Villena-Román, José-Carlos González-Cristóbal, Cristina Moreno-García, and José-Luis Martínez-Fernández.

This paper describes a commercial project which applies the concepts put forward by the Semantic Web in order to improve image search in a website for selling photographs through the Internet. The specific problem addressed here concerns techniques for the semiautomatic creation of thesauri and the normalization of image descriptors from a previous set of labels showing free keywords with partial morphological expansion. The ultimate goal of this project is to improve the customer accessibility to a collection of more than two million photographs. This project has been developed by the Spanish company DAEDALUS-Data, Decisions and Language, S.A. for the Internet website stockphotos.es, of the company Stock Photos S.L.

Keywords: Automatic Classification, Digital Image Library, Information Retrieval, Normalisation Process, Post-Translation, Pretranslation, Subject Hierarchy, Thesaurus.

1 Introduction

The ultimate goal of the Semantic Web is to improve the access to any kind of information that is published in on the Internet. Nowadays, several languages and standards, which are mainly promoted by W3C (World Wide Web Consortium), allow a uniform representation of information, as well as the formalization of inference processes. Both aspects are essential to facilitate the localization of information stored in any digital repository.

Today, these standards and supporting tools make it very easy to adopt an ontology for a particular domain, perhaps with some kind of adaptation, or even to build it one from scratch for that particular domain or a specific application. These are areas which need a limited effort and are, in general, achievable in commercial projects. However, content providers also want to reach the promised land of the Semantic Web from a huge volume of information which is few not highly structured or is totally unstructured. Moreover, as large amounts of money and effort have already been invested in the acquisition or Taylorization of those resources, any action to make profit from them needs to balance the accessibility by customers and economical constraints.

This is the main problem that arises in the case that we present in this article: how to improve the likelihood that a given customer finds the photograph which he/she needs to illustrate a publication or an advertising campaign, in the shortest time, in an archive with several million images. This objective necessarily demands that the images are tagged in the best way to match the user query. But tagging a photograph means specifying the objects that are shown, the environment in which they are located, relationships among them, actions or effects which could be happening at that moment, feelings that are evoked, light, colour range, photographic technique, etc. This work has already been carried out in the past, perhaps with criteria, depth, precision or quality which that are less than optimal. Logically, therefore, investing more money is not an option.

The situation is a digital image archive, tagged with a short title and several keywords from a free (uncontrolled) vocabulary, without diacritics or typographical marks. A stratified sampling was performed at the beginning of the project. Finally, the selected set was formed by with 194,618

Julio Villena-Román graduated as a Telecommunications Engineer from the Higher Technical School of Telecommunications Engineers (ETSIT) of the *Universidad Politécnica de Madrid* (UPM), Spain, in 1997. He was a founder member and is currently the technological director of the Spanish company DAEDALUS. He started his career as a researcher at UPM on an FPI (Research Personnel Training) grant, 1997. He has been lecturing in the Telematic Systems Engineering Department of the *Universidad Carlos III de Madrid* since 2002. He has led research projects in the field of intelligent systems and has authored many international publications. <jvillena@it.uc3m.es>

José-Carlos González-Cristóbal is a Doctor of Telecommunications from the Higher Technical School of Telecommunications Engineers (ETSIT) of the *Universidad Politécnica de Madrid* (UPM), Spain, in 1989. He was a founder member and has been the president of the Spanish company DAEDALUS since its inception in 1998. He has been lecturing in the Telematic Systems Engineering Department (ETSIT – UPM) desde 1985. He has led numerous research and development projects, as part of national and European programmes, or private initiative projects. He has authored a great many scientific and technical publications in various fields related to artificial intelligence, and he has taken part as an organizer and collaborator in national and international conferences. He has represented UPM as the Chairman of the Technical Committee of CITAM (Research Centre for Multimedia Technologies and Applications, AIE), and is also the chairman of the Spanish chapter of the Computer Society (IEEE). <jgonzalez@gsi.dit.upm.es>

Cristina Moreno-García graduated in Technical Computer Engineering from the UNED (*Universidad Nacional de Educación a Distancia*, Spanish National Open University). She has been working in the technical department at the Spanish company DAEDALUS since 2000 where she has carried out important work on web technology projects. <cmoreno@daedalus.es>

José-Luis Martínez-Fernández graduated as a Telecommunications Engineer from the Higher Technical School of Telecommunications Engineering (ETSIT) of the *Universidad Politécnica de Madrid* (UPM), Spain, in 1998. He was a founder member of the Spanish company DAEDALUS and is currently its director of consulting. Between 2000 and 2001 he worked at SGI (Soluciones Globales de Internet), a business unit of the Spanish group GMV Sistemas S.A.. He has been lecturing in the Computer Science Department of the *Universidad Carlos III de Madrid* since 2002. He has led research projects in the field of intelligent systems, a subject on which he has authored many international publications. <jlmferna@inf.uc3m.es>

images, tagged with 1,008,593 terms in titles and 2,917,973 terms in keywords. As we will see, different inflectional lexical forms were frequently included in the same image to increase the possibility of it being found. Also a certain proportion of spelling mistakes was found.

The objective of this article is to illustrate the process of term normalization and, at the same time, the creation of an ad-hoc thesaurus which allows to access to all the available contents in a structured and optimised way. This process is particularly demanding in most projects related to Semantic Web.

Questions that arise under these circumstances are:

1. Is it possible, by semiautomatic means and with acceptable costs, to carry out the generation of an ontology suited to the contents of this collection, the normalization of the keys and the multiclassification of those contents according to the generated ontology?
2. To what extent is this project economically worthwhile?
3. What other tools or investments are necessary so that the customers can find the appropriate images within the new content structure?
4. What is the impact of changes on the costs of cataloguing new collections and what are the repercussions of adopting the new technology on the maintenance costs?
5. And, lastly, what is the return of on this investment?

From here, Sections 3 to 6 of this article are focused on the first question, describing the methodology followed in the project. Section 2 is dedicated to putting this work in context,. It showings the relationship of this work to other works in the area of image annotation for the Semantic Web

and also with other R&D projects in compatible areas done by the Spanish company DAEDALUS. Section 7 is dedicated to presenting some conclusions.

2 Framework

This project is connected to other R&D projects carried out by DAEDALUS in the Information Retrieval field: Omnipaper (Smart Access to European Newspapers, IST-2001-32174) [1][2] and EDDENN (*Extracción de Datos de Documentos con Estructura No Normalizada* – Data Extraction from Documents with Non-Normalized Structure, FIT-350200-2004-33 y 350100-2005-308, in collaboration with IPSA). Moreover, this project benefits from the participation of DAEDALUS in the European Information Retrieval forum, specifically anything that is concerned with multilingual image retrieval [3][4].

The present work is related, although with an approach and goals that are very practical, to research activities in the linguistic annotation for the Semantic Web [5], particularly to the annotation of multimedia objects [6][7]. The application of thesauri and ontologies has been explored, for example, by projects to publish Finnish museum pictures on the Internet [8][9] and images in general [10]. On some occasions, pure textual techniques are combined with others with specific image content handling, as in [11] or in our own experiences in [4].

3 Description of The Image Archive

The information source is the digital image archive of StockPhotos, mainly consisting of digital high-resolution colour photographs, in different formats and of heteroge-

ID	JAP-000401-LAI
Collection	ETNICAS-III
Title	JAPON ASIA ORIENTE ORIENTAL ASIATICO
Keywords	COLOR PAISAJE MUJERES <i>MUJER</i> JAPONESAS <i>JAPONESA</i> RAZAS <i>RAZA</i> ETNIAS <i>ETNIA</i> SOMBRILLAS <i>SOMBRILLA</i> PARAGUAS SENTADAS <i>SENTADA</i> <i>SENTARSE</i> <i>SENTAR</i> RELAJADAS <i>RELAJADA</i> <i>RELAJARSE</i> <i>RELAJAR</i> RELAZ ARENA MIRANDO <i>MIRAR</i> PAISAJE TABLAS <i>TABLA</i> SURF TRANSPORTES <i>TRANSPORTE</i> DEPORTES <i>DEPORTE</i> AGUA MAR <i>OCEANOS</i> OCEANO PLAYA ARQUITECTURA EDIFICIOS <i>EDIFICIO</i> RASCACIELOS RASCACIELO CIUDADES <i>CIUDAD</i> MODERNAS <i>MODERNA</i> ODAIBA JAPON TOKIO ASIA VIAJES <i>VIAJE</i> ATRACCION TURISTICA TURISMO TIEMPO LIBRE OCIO ENTRETENIMIENTO DIVERSION <i>DIVERTIR</i> RECREACION <i>RECREAR</i>

ID	JAP-000401-LAI
Collection	ETHNICS-III
Title	JAPAN ASIA ORIENT ORIENTAL ASIATIC ASIAN
Keywords	COLOUR LANDSCAPE WOMEN <i>WOMAN</i> JAPANESE RACES <i>RACE</i> ETHNIC GROUP PARASOL UMBRELLA SAT SIT RELAXED RELAX SAND LOOKING <i>LOOK</i> LANDSCAPE SURFBOARD SURF TRANSPORTATION SPORTS <i>SPORT</i> WATER SEA <i>OCEANS</i> OCEAN BEACH ARCHITECTURE BUILDINGS <i>BUILDING</i> SKYSCRAPER CITIES <i>CITY</i> MODERN ODAIBA JAPAN TOKYO ASIA TRAVEL ATTRACTION TOURISTIC TOURISM FREE TIME HOBBY ENTERTAINMENT AMUSEMENT LEISURE

Figure 1: Example of Information about An Image. (The English version of the examples is not exactly aligned with the original Spanish one as there are several more words, mainly plurals, in the latter.)

neous subjects, including both copyright and royalty-free images. A stratified sampling based on subjects was performed at the beginning of the project, to finally build a set of more than 100,000 images.

As long well as the photograph, each image in the archive had several structured information fields associated with it. Apart from the image identifier (unique id) and the collection in which the image is included, two other text fields were relevant to this project: the title (short description) of the image and keywords, i.e. additional terms which complement the description of the image. Actually both fields are concatenated and handled as one single field. As in the example in **Figure 1**, fields have are free text (without a controlled vocabulary), in Spanish, and describe the image from some points of view: format, technique, author or agency, etc., apart from the image content (objects in foreground or background, concepts and feelings, number, sex and age of people shown, geographical information, common use synonyms, etc.).

This information is indexed in a database management system, which is exploited with internal tools and also from the company's public website. Visitors may make queries which combine search terms and the usual logical operators (AND, OR and NOT), view the images and their attributes, and may make the purchase.

Keywords are variable-length literals which are words separated by one or more spaces. To simplify the edition editing and search processes, all words are converted to uppercase and accents are eliminated. Among the keywords there are some ambiguous terms (for example *ratón* [mouse], a small animal or a computer peripheral) and also grammatical words (those which have no meaning on their own, such as prepositions or conjunctions, whose main function is to build the syntactic structure of the sentence).

As there is no specific separation between indexing terms (different from the other than space), multiword terms are undifferentiated from the others (for example: *primer plano* [foreground], *tiempo libre* [free time]). The text also includes incorrect words, due to spelling mistakes or typing errors, more frequently than desirable, as it is usual in documentary databases.

Furthermore, due to the highly inflectional nature of the Spanish morphology, additional indexing terms are generated by semiautomatic means to improve the system recall by increasing the number of results. These terms are the inflectional forms (nominal and/or verbal) corresponding to the original terms, and some of their synonyms. For example:

- Nominal inflection (singular ↔ plural): *mujeres* → *mujer* [women → woman]

- Verbal inflection: (participle ↔ infinitive): *sentadas* → *sentada/sentarse/sentar* [*sat* (plural participle) → *sat* (singular participle)/*sits* (reflexive)/*sit*]

- Synonyms: *sombrilla* → *paraguas* [*parasol* → *umbrella*]

The objective is to be able to find a given image, independently of the specific word forms which the user includes in his/her query, for example, *edificio/edificios* [building/buildings], *relajadas/relajar* [relaxed/relax] or *sombrilla/paraguas* [*parasol/umbrella*]. It is interesting to note that, in the case of nominal words, gender inflectional forms are not included (masculine → feminine) to avoid misleading the user when, for instance, he/she looks for *niña bailando* [girl dancing] and results include images with *niños* [boys].

Due to this automatic process, a well-known problem is that, in multiword terms, original and additional words are intercalated and mixed up (*ciudades modernas* → *ciudades ciudad modernas moderna* [modern cities → modern(plural)

modern(singular) cities city]), as no distinction is possible between single and multiword terms. Another problem is that some terms may be duplicated, depending on the order in which the expansion has been done.

Many descriptions have been machine-translated from other languages (English, French, German), with variable quality. Moreover, depending on their origin, some descriptions include Spanish idiomatic expressions or regional inexpressions from Spain or Latin America, which may make queries more difficult for different Spanish speakers.

In short, due to the diverse origins of the resources and despite the exhaustive pre-processing tasks, descriptions are very heterogeneous and their quality changes varies among different collections or even among different images in the same collection.

4 Thesaurus Construction

The objective of the first phase of the project was the construction of a conceptual classification thesaurus for the digital image archive, semantically representative of its contents (archive coverage). The final result of the process would be a catalogue with nodes distributed in a hierarchical tree structure, where each node represents a concept (categories), and may include one or more descriptive terms related to that concept and, in some cases, other nodes which would represent more specific concepts (subcategories). For example, when the concept is a "place", descriptive terms would be "geographic references" (city names, monuments, rivers or mountains, including synonyms) and more specific concepts would be "smaller territorial divisions" (such as countries or states). From this point of view, the thesaurus was a semantic net in which nodes (concepts and descriptors) are related among them with the relationships *describes/is_described_by* (descriptor ↔ concept) and *more_general/more_specific* (concept ↔ subconcept).

After an initial research on existing classification hierarchies in the area, none of which was perfectly suited to our purposes, our methodology consisted of the combination of the analysis of the thematic contents of the archive, the know-how in the company and the behaviour of website visitors, in an iterative process with a spiral life cycle.

The final classification hierarchy didn't did not focus on a general purpose but was rather pragmatically designed. Our client was looking for a thesaurus which was closely adapted to their archive contents, and insisted on priority being given to more frequently used categories with less emphasis on (or directly omitting) those categories with few images. Therefore, the tree is not balanced. The maximum depth corresponds to categories *alimentación* [food] and *naturaleza* [nature] (5 levels), compared to category *monumentos* [monuments] (2 levels) in which one lower specification level was enough. In its final version, the thesaurus includes 34 root categories and 276 categories in all.

XML (eXtensible Markup Language) was adopted for the thesaurus representation from the beginning, as it is considered to be the most suitable for describing tree-structured data, as in our case. Moreover, XML is platform-independent, permits internationalization as it is fully Unicode compatible, its computational management is quite easy and, as it is a text-based format, it is possible to read and/or edit XML documents with standard well-known editing tools, if necessary. The main drawback was the increase in the size of the thesaurus due to the tag and syntax overhead, but its impact was not considered to be relevant.

Finally, the data structure of the thesaurus is defined with the DTD (Document Type Definition) shown in **Figure 2**.

```

<!ELEMENT thesaurus (concepts)>
<!ELEMENT concepts (concept+)>
  <!ELEMENT      concept      (name,
descriptors?, concepts?)>
  <!ELEMENT descriptors (descriptor+)>
  <!ELEMENT name #PCDATA>
  <!ELEMENT descriptor #PCDATA>

```

Figure 2: Thesaurus DTD.

Regarding the descriptors, the design guidelines for the team of linguists were that descriptors ought to be, whenever possible, word lemmas. In high inflectional languages such as Spanish, lemmas are the paradigmatic forms that represent the whole set of inflectional forms that can be obtained with a morphological process, that is, the representatives of the different variants of the same word. Usually the lemma is the masculine-singular form for nominal words (nouns, adjectives, determinatives and pronouns) – when that form exists; if not, feminine, or plural –, the infinitive for verbal words (verbs) and the same word for the rest (prepositions, conjunctions, etc.). Lemmas allow performing performance of any linguistic operation independently of the specific variant of the word.

The thesaurus includes both single and multiword terms (in this case, different words are concatenated with an underscore). To distinguish among ambiguous meanings of different terms, compound descriptors in the form *term/meaning* are used, for instance: *cabo/geografía* [*cape/geography*] and *cabo/militar* [*corporal/army*]. The thesaurus includes over 7,000 descriptors in its final version.

5 Normalization Process

The objective of the next stage of the project was to establish a matching between each keyword in the image

descriptions and the corresponding descriptor(s) in the thesaurus. This stage was named the *normalization process*, i.e., "categorize, adjust to a model, rule or norm" (according to the *Real Academia Española*, Royal Academy of the Spanish Language), referring to the process of transforming words in a free (uncontrolled vocabulary) text into terms of a restricted or controlled vocabulary.

Obviously, the definition of the normalization process should take into account the specific features and problems in both the digital archive and the thesaurus. In addition, it was essential to build a robust system in order to anticipate images which will be incorporated into the archive in the future. Therefore, a functional requirement was that any linguistic resource included in the system had to be easily reconfigurable (modified, expanded) by non experts, to simplify system maintenance and evolution.

Finally the normalization process was designed in two cycles (pre-cycle and post-cycle), each one again with two stages (translation and classification), applied in cascade. The final goal is was to classify the original keywords in the image description into two sets: normalized keywords and the remaining keywords. The whole normalization process is described in the next sections.

5.1 Stage 1a. Pretranslation

The first stage of the normalization process, which is called "initial translation" or *pretranslation*, is used to make simple transformations (*translations*) to the image keywords. The goal is to prepare the description in the original text before an initial keyword search in the thesaurus (the second stage). The input for this stage is a text field with the image description, and the output is the transformed text (*translated*).

The process is internally based on the *pretranslation table*, which contains a list of terms with their corresponding translation. Actually this table is a text file with two columns in the format: *original_term \t translated_term*

The first column contains the original term (both single terms and multiword expressions) and the second column, separated by tab, indicates the word or expression in to which

```

<?xml_version="1.0"?>
<!DOCTYPE thesaurus SYSTEM "thesaurus.dtd">
<thesaurus>
  <concepts>
    <concept>
      <name>AGRICULTURE</name>
      <descriptors>
        <descriptor>greenhouse</descriptor>
        ...
      </descriptors>
    <concepts>
      <concept>
        <name>CROPS</name>
        ...
      </concept>
      ...
    </concepts>
  </concept>
  ...
</concepts>
</thesaurus>

```

Figure 3: Extract of The Thesaurus (Translated from Spanish).

<i>viviendas</i> → <i>vivienda</i> [<i>housing</i> → <i>flat</i>]	The descriptor <i>bloque de viviendas</i> (block of flats) is included in the thesaurus, so it is not appropriate to apply the transformation <i>viviendas</i> → <i>vivienda</i> in pretranslation before searching in the thesaurus (it would be wrongly transformed into <i>bloque de viviendas</i> [<i>block of flats</i>])→ <i>bloque de vivienda</i> [<i>block of flat</i>])
<i>estados</i> → <i>estado</i> [<i>states</i> → <i>state</i>]	<i>Estados Unidos</i> → <i>estado Unidos</i>
<i>montañas</i> → <i>montaña</i> [<i>mountains</i> → <i>mountain</i>]	<i>Montañas Rocosas</i> → <i>montaña Rocosas</i>
<i>espera</i> → <i>esperar</i> [<i>wait</i> → <i>to wait</i>]	<i>Sala de espera</i> → <i>sala de esperar</i> [<i>waiting room</i>]

Figure 4: Preventing Mistakes when Transforming Words.

it is transformed. Each line contains one entry. The priority order in for translation descends from the first entry (highest priority) to the last one (lowest priority). No checking is performed when applying the transformation, i.e., if any of the included terms in the pretranslation table is found, it is substituted with the translated term without further verification.

Typically this stage is used to lemmatize the keywords, transforming the inflectional forms in the descriptions into their corresponding lemmas, which are, eventually, the ones that are included as descriptors in the thesaurus. Because of this, most transformations concern gender, number and verb forms, including, again, single or multiword terms, for example:

medicamentos→*medicamento* [*medicines*→*medicine*]
farmacias→*farmacia* [*pharmacies*→*pharmacy*]
cajas registradoras→*caja registradora* [*cash registers*→*cash register*]

This stage also includes transformations for synonyms and other derivative schemas (such as diminutives), multiword expressions and incorrect or partially completed words:

agenda electrónica?pda [*personal digital assistant*→*pda*]
osito→*oso* [*little bear*→*bear*]
centros centro comerciales comercial→*centro comercial*
[*shoppings(plural)* *shopping(singular)* *centres centre*→*shopping centre*]
chirimolla→*chirimoya* [*kustard aple*→*custard apple*]
mercado valores→*mercado de valores* [*stock exchange*]

Moreover, pretranslation may be used to expand the original image descriptions with missing terms. For example:

Sagrada Familia→*Sagrada Familia arte catedral*
turismo estilo gótico [*Sagrada Familia art cathedral tourism style gothic*]

It is very important to remark that it is essential to apply the common sense when adding new entries in the pretranslation table so as not to include transformations of words that are part of a multiword descriptor in the thesaurus. For example **Figure 4**. The pretranslation process makes an automatically verification verifiesof entries in the table which potentially could cause trouble. In its final version, this table includes almost 6,100 entries.

5.2 Stage 1b. Preclassification

The input to this stage is the output of the previous translation process. Then, keywords are searched in the thesaurus. If a term (single or multiword) coincides with a thesaurus descriptor, it is extracted from the input text and added to the normalized keyword set. The final output of this process is a string with the normalized keywords (separated by spaces) and another string with the remaining keywords which are not found in the thesaurus. The search is performed from left to right, choosing the longest valid descriptor (for instance: *silla de cuero*→*silla_de_cuero* [*leather-chair*] and not *silla*→*silla* and *cuero*→*cuero*).

This stage is called *preclassification* because a first separation between normalized and non-normalized keywords is done (later, a second similar stage is executed).

5.3 Stage 2a. Post-translation

The *post-translation* stage (so-called because it is executed after the first classification) constitutes the beginning of a new translation/classification cycle, and its objective is to apply more complex transformation functions on the remaining keywords which have not been normalized in the previous stage and are received as input. Unlike the pretranslation stage, and additionally, transformation rules allow specification of contexts in which the activation of the rule is conditional depending on the presence or absence of that context.

The process is based on the *post-translation table*, whose format is similar to that of the pretranslation table for simple transformation rules, whereas, in conditional transformations, is the following:

original_term \t *translated_term* \t [*conditionYes*] \t [*conditionNo*]

The first two columns are analogous to the ones in simple transformations. The third column (*conditionYes*) contains a word list, separated by commas, at least one of which must be present in the text so that the rule is applicable. In a similar way, the fourth column (*conditionNo*) contains a list of words, none of which must appear in the text to make the transformation effective. These last two columns are optional.

For example:

cabo *cabo/geografía* *mar,geografía*
ejército,militar
cabo *cabo/militar* *militar* *mar*

It means that *cabo*, that means corporal or cape in Spanish, is transformed into *cabo/geografía* [cape/geography] if the image description contains *mar* [sea] or *geografía* [geography] and contains neither *ejército* [army] nor *militar*[soldier], and is transformed into *cabo/militar* [corporal/army] if the word *militar* is present and *mar* [sea] is not present.

Another way to express the same idea could be:

cabo cabo/geografía mar,geografía
cabo cabo/militar militar,ejército

or even:

cabo cabo/geografía militar
cabo cabo/militar mar,geografía

If a default rule had been needed, just in case none of the previous rules was applicable, it had to be included as a simple transformation rule before the others. For example:

cabo cabo/militar

Like pretranslation, post-translation may be used to expand the image descriptions with additional terms, even with conditional decisions, such as:

rastro rastro domingo Madrid
rastro rastro mercado Madrid
 [rastro=flea market, domingo=Sunday, mercado=market]

In the final version, the post-translation table has more than 2,700 entries.

5.4 Stage 2b. Post-classification

The fourth and last final stage of the normalization process performs a new classification of the keywords after the post-translation stage, similar to the first classification. Again, the output of the stage is a string with the normalized keywords (space separated) and another string with the remaining keywords that are not present in the thesaurus.

6 Example of Application

The whole process is illustrated with a real example:

Initial description

deporte ocio campo de entrenamiento tiempo libre retrato grupo de niños niño corriendo jugar fútbol exterior balón pelota

[*sport leisure_time training camp free time portrait group of children child running run playing play soccer exterior football ball*]

Pretranslation

corriendo→correr
jugando→jugar

deporte ocio campo de entrenamiento tiempo libre retrato grupo de niños niño correr correr jugar jugar fútbol exterior balón pelota

[*running→run*]

[*playing→play*]

[*sport leisure_time training camp free time portrait group of children child run run play play soccer exterior football ball*]

Preclassification

Normalized keywords:

deporte ocio entrenamiento tiempo_libre retrato

grupo_de_niños niño correr jugar fútbol exterior balón pelota
[sport leisure_time training free_time portrait group_of_children child run play soccer exterior football ball]

Remaining (non-normalized) keywords:

campo de
[camp of]

Post-translation

campo→campo/deporte deporte,entrenamiento
campo→campo/paisaje

campo/deporte de

[*camp→campo/sport sport,training*]

[*camp→campo/landscape*]

[*camp/sport de*]

Post-classification

Normalized keywords (all):

deporte ocio entrenamiento tiempo_libre retrato grupo_de_niños niño correr jugar fútbol exterior balón pelota campo/deporte
[sport leisure_time training free_time portrait group_of_children child run play soccer exterior football ball camp/sport]

Remaining (non-normalized) keywords:

de
[of]

Finally, a thematic image classification is given simply by locating the normalized keywords in the thesaurus.

Image classification (in Spanish and English):

11. DEPORTES (ENTRENAMIENTO, DEPORTE, CAMPO DEPORTE)

16.5.1. EXTERIOR (EXTERIOR)

16.5. UBICACION

16. FORMALIDADES_DE_LA_FOTO

16.4. TECNICA (RETRATO)

16. FORMALIDADES_DE_LA_FOTO

28.1. CONCEPTOS (OCIO, TIEMPO_LIBRE)

28. NUMEROS, MESES, SENTIMIENTOS_Y_CONCEPTOS

29.4. OBJETOS (BALON, PELOTA)

29. OBJETOS,_COLORES,_FORMAS_Y_MATERIALES

11.6. MODALIDADES_DEPORATIVAS (FUTBOL)

11. DEPORTES

1. ACCIONES_Y_EFECTOS (CORRER, JUGAR)

30. OCIO_Y_TIEMPO_LIBRE (OCIO)

18.5.2. NIÑO (GRUPO_DE_NIÑOS, NIÑO)

18.5. 3-13_AÑOS

18. GENTE

11. SPORTS (TRAINING, SPORT, CAMP/SPORT)

16.5.1. EXTERIOR (EXTERIOR)

16.5. LOCATION

16. PHOTO_PROPERTIES

16.4. TECHNIQUE (PORTRAIT)

16. PHOTO_PROPERTIES

28.1. CONCEPTS (LEISURE_TIME, FREE_TIME)

28. NUMBERS,_MONTHS,_FEELINGS_&_CONCEPTS

29.4. OBJECTS (FOOTBALL, BALL)

29. *OBJECTS, COLOURS, FORMS & MATERIALS*

11.6. *SPORT DISCIPLINES (SOCCER)*

11. *SPORTS*

1. *ACTIONS & EFFECTS (RUN, PLAY)*

30. *LEISURE & FREE TIME (LEISURE TIME)*

18.5.2. *CHILD (GROUP OF CHILDREN, CHILD)*

18.5. *3-13_YEARS_OLD*

18. *PEOPLE*

7 Conclusions

Any engineering project is always conditioned by constraints on available resources and the context in which the solution is going to be applied. It is necessary to make this statement because, in the application domain of the Semantic Web, like in many others, sometimes it is possible to satisfy the client's requirements although the solution doesn't does not fully reach the current limits of technology limits. In our case, there were some determining factors which led us to propose a search system based on an ad-hoc thesaurus using a normalized set of descriptors, instead of the development of a formal complete ontology based on RDF/OWL. The final implementation was done in XML so that the representation format did not cause much disturbance in our client's technological technical environment. The development of the complete system, including all the auxiliary tools for the cataloguing of new images, etc., took 4 person-months of effort.

The solution incorporated available tools from DAEDALUS for performing the lemmatization and Part-Of-Speech tagging in Spanish (STILUS[®] Core library)¹, spell checking (K-Site[®] Fuzzy library) and semantic expansion (STILUS[®] Sem). The integration of these tools allowed us to reach a high degree of automation in the normalization processes.

More important, the final system allows the users to make queries which are structured according to the usual criteria in the image archive market. All those linguistic make it possible to transform a query, expressed either in natural language or with keywords, into normalized terms used in the indexes, with an efficient, fully automatic process of analysis, checking and semantic expansion.

Last, but not least, the approach adopted in this project let us meet, efficiently and economically, the client requirements without imposing substantial changes in their technological production environment. The return on investment may be quantified with metrics relative to the average mean time to execute a query, the decrease in non-completion rate (rate of queries without results), or the increase in customer confidence or loyalty, all of which have a direct impact in on business figures.

Acknowledgements

We specially thank StockPhotos for their permission to disseminate the results of this project. StockPhotos is an agency associated to with the LatinStock gGroup. LatinStock is an organization currently established in Argentina, Brazil, Chile, Mexico, Spain, Colombia, Venezuela, Peru, Uruguay and Costa Rica, whose main field of activity is to supply the publishing, advertising and television industries in Latin America and Spain with image contents.

References

- [1] José C. González, Julio Villena, Francisco Bueno, Ana M. García Serrano y Paloma Martínez. OMNIPAPER: Smart Access to European Newspapers. In Actas de la Conferencia de la Sociedad Española para el Procesamiento de Lenguaje Natural, SEPLN 2002, Valladolid, 2002.
- [2] T. Pereira and A.A. Baptista. A The OmniPaper Metadata RDF/XML Prototype Implementation. In Proceedings of the 7th International Conference on Electronic Publishing, Portugal, 2003.
- [3] Julio Villena, José L. Martínez, Jorge Fombella, Ana G. Serrano, Alberto Ruiz, Paloma Martínez, José M. Goñi and José C. González. Image Retrieval: The MIRACLE Approach. In "Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003". Carol Peters et al. (editores). Lecture Notes in Computer Science Vol.3237, pp. 621–630. Springer-Verlag, 2004.
- [4] J.L. Martínez-Fernández, A. García-Serrano, J. Villena, V. Méndez-Sáenz. MIRACLE Approach to Image CLEF 2004: Merging Textual and Content-Based Image Retrieval. In "Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004". Lecture Notes in Computer Science Vol.3491, pp. 699–708. Springer-Verlag, 2005.
- [5] Paul Buitelaar and Thierry Declerck. Linguistic Annotation for the Semantic Web. In Siegfried Handschuh, Steffen Staab (eds.). Annotation for the Semantic Web, IOS Press, 2003.
- [6] Joost Geurts, Jacco van Ossenbruggen and Lynda Hardman Requirements for Practical Multimedia Annotation. In Proceedings of the Workshop "Multimedia and the Semantic Web", 2nd European Semantic Web Conference, Heraklion, Crete, 2005.
- [7] Jung-ran Park. Semantic Interoperability across Digital Image Collections: a Pilot Study on Metadata Mapping. In Lecture Notes in Computer Science, Vol. 3237, pp. 621-630. Springer-Verlag 2004.
- [8] Eero Hyvönen, Avril Styrman and Samppa Saarela. Ontology-Based Image Retrieval. Report 2002-03 in HIIT Publications, pages 43-45. Helsinki Institute for Information Technology (HIIT), Helsinki, Finland, 2002.
- [9] Eero Hyvönen, Mirva Salminen, Miikka Junnila, Suvi Kettula. A Content Creation Process for the Semantic Web. In Proceedings of OntoLex 2004: Ontologies and Lexical Resources in Distributed Environments, Lisbon, Portugal, 2004.
- [10] L. Hollink, G. Schreiber, J. Wielemaker, and B. Wielinga. Semantic annotation of image collections. In Proceedings of the Workshop on Knowledge Markup and Semantic Annotation, K-CAP'03, Florida, 2003.
- [11] Shuqiang Jiang, Tiejun Huang, Wen Gao. An Ontology-based Approach to Retrieve Digitized Art Images, pp. 131-137, IEEE/WIC/ACM International Conference on Web Intelligence (WI'04), 2004.

¹ STILUS y K-Site are trade marks of DAEDALUS-Data, Decisions and Language, S.A.